

A Generalized Kruskal–Wallis Test Incorporating Group Uncertainty with Application to Genetic Association Studies

Elif F. Acar^{1,2} and Lei Sun^{3,1}

¹ Department of Statistics, University of Toronto

² Department of Mathematics and Statistics, McGill University (current affiliation)

³ Dalla Lana School of Public Health, University of Toronto

Abstract

Motivated by genetic association studies of SNPs with genotype uncertainty, we propose a generalization of the Kruskal–Wallis test that incorporates group uncertainty when comparing k samples. The extended test statistic is based on probability-weighted rank-sums and follows an asymptotic chi-square distribution with $k - 1$ degrees of freedom under the null hypothesis. Simulation studies confirm the validity and robustness of the proposed test in finite samples. Application to a genome-wide association study of type 1 diabetic complications further demonstrates the utilities of this generalized Kruskal–Wallis test for studies with group uncertainty.

Keywords: *Genome-wide association studies; Imputation; k -sample problem; Misclassification; Next-generation sequencing; Non-parametric test; Probabilistic data; Rank.*

1 Introduction

The seminal work by Kruskal and Wallis (1952) provided us a robust rank-based test for the k -sample problem, complementing the parametric approaches such as the one-way analysis of variance (ANOVA). In the classical k -sample problem, data are well classified into different categories or groups. However, in many current scientific studies the categorical variables are not necessarily deterministic, and the uncertainties are quantitatively expressed by probability distributions over attributes. Such classification problems often arise in biomedical and bioinformatics applications where data mining techniques and classification algorithms are used to obtain class membership probabilities.

A particular motivating example of this work is the genetic association study of single nucleotide polymorphisms (SNPs) for which the genotype group assignments are often known with ambiguity. The data uncertainties at these SNPs are typically represented by genotype probabilities obtained from various genotype calling algorithms (e.g. [Carvalho et al., 2010](#)) or imputation algorithms (e.g. [Li et al., 2009](#)). Table 1 provides a hypothetical illustration. In such cases, a number of parametric remedies have been proposed, including the popular dosage approach, the weighted regression method ([Aulchenko et al., 2010](#)) and likelihood-based score tests ([Schaid et al., 2002](#)). Although these parametric approaches are satisfactory in many applications, investigators often seek complimentary evidence provided by robust non-parametric alternatives, safeguarding their statistical analyses against potential model misspecifications. Therefore, it is of both theoretical and practical importance to generalize the Kruskal–Wallis test so that it is applicable to the k -sample problem but with group uncertainty.

To formulate the testing problem, let Y be a continuous response variable and G a categorical variable with k distinct attributes. For instance, in genetic association studies, Y denotes the phenotype of interest (e.g. blood pressure or glucose level) and G is the genotype variable at a particular SNP with $k = 3$. The three categories for a SNP represent if an individual’s genotype at this SNP contain 0, 1 or 2 copies of the minor allele (one of the two alleles with population frequency < 0.5). The goal of the association analysis between Y and G is to determine if the phenotype Y values differ between individuals with different genotype G values. This can be achieved,

Table 1. An illustration of SNP genotype probabilities

Individual	Genotype			hard call	soft call
	0	1	2		
1	0.925	0.045	0.030	0	0.105
2	0.156	0.102	0.742	2	1.586
\vdots	\vdots	\vdots	\vdots		
N	0.375	0.410	0.215	1	0.840

for example, by regressing Y on G and other relevant covariates. Hence, the classic linear model

$$Y_j = \mu + \beta G_j + \varepsilon_j, \quad j = 1, 2, \dots, N, \quad (1)$$

with $\varepsilon_j \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, provides a basis for most association analysis or group comparisons. However, results from the robust Kruskal–Wallis test are often obtained as preliminary or complimentary evidence.

In practice, what is available to us may not be the true G group value, but rather probabilistic data of G , i.e. the vector of group probabilities $\mathbf{p}_j = (p_{1j}, p_{2j}, \dots, p_{kj})$, where $p_{ij} = P(G_j = i)$ for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, N$, with $\sum_{i=1}^k p_{ij} = 1$. In genetic association studies of SNPs, depending on the experiment used by each application, the genotype of a SNP may be inferred via classification algorithms (e.g. Birdseed, (Korn et al., 2008)), imputed via imputation algorithms, (e.g. TUNA (Nicolae, 2006), MaCH (Li et al., 2006) and Impute (Marchini et al., 2007)), or derived from next generation sequencing calling algorithms (e.g. SYZYGY (Calvo et al., 2010) and SNVer (Wei et al., 2011)). In each case, an individual’s genotype is most likely associated with some level of uncertainty and the probability that the true genotype belongs to each of the three genotype groups, $p_{ij} = P(G_j = i)$ for $i = 0, 1, 2$ is provided for individual j , as illustrated in Table 1.

In the presence of genotype group uncertainty, the best-guess approach (also known as the hard-call approach) bypasses the problem of incomplete group information by using the most probable group, $\tilde{G}_j = \{i : p_{ij} = \max(p_{0j}, p_{1j}, p_{2j})\}$, in place of G_j in (1). Depending on whether \tilde{G}_j are considered ordinal or categori-

cal, a t -test ($df = 1$) or an ANOVA F -test ($df = 2$) can be performed, so is the Kruskal–Wallis test ($df = 2$). We refer to these tests as Best-Guess Linear Model (BG-LM), Best-Guess ANOVA (BG-ANOVA) and Best-Guess Kruskal–Wallis (BG-KW), respectively (Table 2). Although convenient, this best-guess approach fails to fully utilize the information in the group probabilities.

An alternative method is the dosage approach (also known as the expectation-substitution or soft-call approach). In this approach, each G_j in (1) is substituted by its expectation $\bar{G}_j = p_{1j} + 2 \times p_{2j}$. The association evidence is then assessed, for example, by a t -test ($df = 1$) from the regression of Y on \bar{G} . We refer to this test as the dosage test. The main disadvantage of this approach is that it does not allow G to be of categorical nature, and it constrains the relationship between Y and G to an additive model. There are several other model-based methods that incorporate group probabilities (e.g. Marchini et al., 2007; Lin et al., 2008; Kutalik et al., 2010; Aulchenko et al., 2010; Schaid et al., 2002), however, model-free counterparts such as the Kruskal–Wallis test has not been proposed.

The remaining paper is organized as follows. In Section 2, we describe the construction of the generalized Kruskal–Wallis test statistic based on intuitive probability-weighted rank-sums, and provide theoretical justifications of its asymptotic chi-square distribution. We show that the original Kruskal–Wallis test is a special case of the proposed test and discuss how to handle tied observations and possible variations in probabilistic data. Section 3 contains simulation studies to evaluate the finite sample performance of the proposed test at different levels of group uncertainty. Section 4 applies the method to data from a genome-wide association study of complications in type 1 diabetic patients. Section 5 concludes with additional discussions. Additional simulation results are provided in the Supplemental Material.

2 The Generalized Kruskal–Wallis Test

2.1 Notation and the Original Kruskal–Wallis Test

Consider a random sample of size N from a large population consisting of $k \geq 2$ disjoint groups or categories, with population proportions π_i , $i = 1, \dots, k$, and $\sum_{i=1}^k \pi_i = 1$. Denote by G the categorical variable taking values on $\mathcal{G} = \{1, 2, \dots, k\}$

and suppose that each category is adequately represented in the sample. Of interest is to compare the k groups, of sizes n_1, \dots, n_k with $\sum_{i=1}^k n_i = N$, based on a continuous response variable Y . Formally, letting the distribution function of Y over the group i be of the form $F_i(y) = F(y - \theta_i)$, we wish to test

$$H_0 : \theta_1 = \theta_2 = \dots = \theta_k \quad \text{against} \quad H_A : \text{not all } \theta_i \text{'s are equal.} \quad (2)$$

When the precise category assignment of G is available, the Kruskal–Wallis test for (2) is performed by ranking all the observations together and comparing the sum of the ranks for each group. Let r_j be the rank of Y_j in the overall sample and define the indicator variable $Z_{ij} = \mathbf{1}(G_j = i)$ for the group membership, the Kruskal–Wallis test statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1), \quad (3)$$

where $n_i = \sum_{j=1}^N Z_{ij}$ and

$$R_i = \sum_{j=1}^N Z_{ij} r_j.$$

Under the null hypothesis of (2), H follows an asymptotic chi-square distribution with $k - 1$ degrees of freedom (Kruskal, 1952; Kruskal and Wallis, 1952).

2.2 The generalized Kruskal–Wallis Test

Suppose available to are not G but probabilistic data of G , $\mathbf{p}_j = (p_{1j}, p_{2j}, \dots, p_{kj})$, where $p_{ij} = P(G_j = i)$ for $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, N$, with $\sum_{i=1}^k p_{ij} = 1$. In this case, it is intuitive to consider the weighted rank-sum

$$R_i^* = \sum_{j=1}^N p_{ij} r_j \quad (4)$$

for each group as a basis of comparison.

However, direct replacement of R_i with R_i^* in the original Kruskal–Wallis test statistic (3) does not lead to a properly calibrated test statistic. Below we describe the construction of the generalized Kruskal–Wallis test, based on this appealing weighted rank-sum R_i^* , that has an asymptotic chis-square distribution with $k - 1$ degrees of freedom.

We make the following assumptions.

(A1) \mathbf{p}_j 's are independent of Y_j 's.

(A2) $\bar{p}_i = \sum_{j=1}^N p_{ij}/N \rightarrow \pi_i$, as $N \rightarrow \infty$, with $0 < \pi_i < 1$.

(A3) $\sum_{j=1}^N (p_{ij} - \bar{p}_i)^2/N \rightarrow \nu_i > 0$, as $N \rightarrow \infty$.

(A4)

$$\frac{\sum_{j=1}^N (p_{ij} - \bar{p}_i)(p_{i'j} - \bar{p}_{i'})}{\sqrt{\sum_{j=1}^N (p_{ij} - \bar{p}_i)^2 \sum_{j=1}^N (p_{i'j} - \bar{p}_{i'})^2}} \rightarrow \rho_{ii'} \text{ as } N \rightarrow \infty.$$

The independence assumption (A1) is standard in statistical analyses of explanatory/response data and is reasonable in practice because, for instance, most imputation algorithms do not consider the phenotype data in the classification of the genotype variable ([Marchini and Howie, 2010](#)). The assumption (A2) ensures that $\sum_{j=1}^N p_{ij}$ provides a reasonable approximation for n_i and that the relative group sizes are convergent. Assumptions (A3) and (A4) are required for the (joint) asymptotic normality of the linear rank statistic R_i^* .

If the groups are indeed from an identical population, the r_j 's can be viewed as a random sample drawn without replacement from the first N integers. Thus, $E(r_j) = (N+1)/2$, $\text{Var}(r_j) = (N^2-1)/12$ and $\text{Cov}(r_j, r_{j'}) = -(N+1)/12$, for $j \neq j'$ under the null hypothesis of (2). The conditional mean and the conditional variance of R_i^* given the group probabilities p_{ij} can then be derived as

$$E(R_i^*) = \tilde{\mu}_i = \frac{N+1}{2} \sum_{j=1}^N p_{ij} \quad \text{and} \quad \text{Var}(R_i^*) = \tilde{\sigma}_i^2 = \frac{N(N+1)}{12} \sum_{j=1}^N (p_{ij} - \bar{p}_i)^2,$$

respectively. An important distinction between our approach and that of [Kruskal and Wallis \(1952\)](#) is that we consider permutations of the first N integers rather than finite-sampling, hence no finite sample correction is required in our derivations.

2.2.1 The case with two samples

The following result is due to the asymptotic theory of linear rank statistics ([Wald and Wolfowitz, 1944](#); [Hájek et al., 1999](#)) and governs our test construction.

Theorem 1. *Under the null hypothesis of (2) and assumptions (A1)-(A3), the limiting distribution of*

$$L_N = \frac{R_i^* - (N+1)/2 \sum_{j=1}^N p_{ij}}{\tilde{\sigma}_i}$$

is standard normal with mean 0 and variance 1.

Proof. It suffices to show that the sequence $(p_{i1}, p_{i2}, \dots, p_{iN})$ satisfies the W condition of the Wald–Wolfowitz Theorem (Wald and Wolfowitz, 1944), i.e.

$$\frac{N^{-1} \sum_{j=1}^N (p_{ij} - \bar{p}_i)^r}{\left\{ N^{-1} \sum_{j=1}^N (p_{ij} - \bar{p}_i)^2 \right\}^{r/2}} = O(1), \quad r = 3, 4, \dots$$

Since $0 \leq p_{ij} \leq 1$, with $\bar{p}_i \rightarrow \pi_i$, the central sample moments of the sequence $(p_{i1}, p_{i2}, \dots, p_{iN})$ are finite:

$$\frac{1}{N} \sum_{j=1}^N (p_{ij} - \bar{p}_i)^r = O(1), \quad \text{for } r = 3, 4, \dots$$

Assumption (A3) ensures a non-zero variance for $(p_{i1}, p_{i2}, \dots, p_{iN})$ and hence the W condition holds. For a detailed proof of the asymptotic normality of L_N see, for example, Theorem 6.1 of Fraser (1957). \square

The generalized Kruskal–Wallis test for the two-sample problem takes the form

$$H^* = \left(R_i^* - \frac{N+1}{2} \sum_{j=1}^N p_{ij} \right)^2 / \tilde{\sigma}_i^2, \quad (5)$$

where $i = 1$ or 2 . By Theorem 1, H^* has an asymptotic $\chi^2(1)$ distribution under the null hypothesis. Note that it is sufficient to consider only one of the R_i^* 's in H^* as

$$\frac{R_1^* - (N+1)/2 \sum_{j=1}^N p_{1j}}{\tilde{\sigma}_1} = - \frac{R_2^* - (N+1)/2 \sum_{j=1}^N p_{2j}}{\tilde{\sigma}_2},$$

which can be easily verified using $p_{1j} = 1 - p_{2j}$, for $j = 1, \dots, N$.

2.2.2 The case with three samples

For the three-sample problem, we shall consider the joint distribution of any two R_i^* and $R_{i'}^*$. The covariance between R_i^* and $R_{i'}^*$ can be calculated as

$$\text{Cov}(R_i^*, R_{i'}^*) = \frac{N(N+1)}{12} \sum_{j=1}^N (p_{ij} - \bar{p}_i)(p_{i'j} - \bar{p}_{i'}),$$

which yields the correlation

$$\tilde{\rho}_{ii'} = \frac{\sum_{j=1}^N (p_{ij} - \bar{p}_i)(p_{i'j} - \bar{p}_{i'})}{\sqrt{\sum_{j=1}^N (p_{ij} - \bar{p}_i)^2 \sum_{j=1}^N (p_{i'j} - \bar{p}_{i'})^2}}.$$

Theorem 2. *Under the null hypothesis of (2) and assumptions (A1) – (A4), the limiting distribution of*

$$L_N = \frac{R_i^* - (N+1)/2 \sum_{j=1}^N p_{ij}}{\tilde{\sigma}_i} \quad \text{and} \quad L'_N = \frac{R_{i'}^* - (N+1)/2 \sum_{j=1}^N p_{i'j}}{\tilde{\sigma}_{i'}}$$

is bivariate normal with means 0, variance 1 and correlation $\rho_{ii'}$.

Proof. Since the sequences $(p_{i1}, p_{i2}, \dots, p_{iN})$ and $(p_{i'1}, p_{i'2}, \dots, p_{i'N})$ satisfy the W condition, the result directly follows from Theorem 6.3 of Fraser (1957). \square

Similar to Kruskal and Wallis (1952), the generalized Kruskal–Wallis test can be formed by considering the exponent of the bivariate normal distribution of Theorem 2, multiplied by -2,

$$H^* = \frac{1}{1 - \tilde{\rho}_{ii'}^2} \left[\left(\frac{R_i^* - (N+1)/2 \sum_{j=1}^N p_{ij}}{\tilde{\sigma}_i} \right)^2 + \left(\frac{R_{i'}^* - (N+1)/2 \sum_{j=1}^N p_{i'j}}{\tilde{\sigma}_{i'}} \right)^2 - 2\tilde{\rho}_{ii'} \left(\frac{R_i^* - (N+1)/2 \sum_{j=1}^N p_{ij}}{\tilde{\sigma}_i} \right) \left(\frac{R_{i'}^* - (N+1)/2 \sum_{j=1}^N p_{i'j}}{\tilde{\sigma}_{i'}} \right) \right], \quad (6)$$

which is asymptotically distributed as $\chi^2(2)$ under the null hypothesis.

An algebraic simplification of H^* is difficult because of the p_{ij} 's. However, as in the two-sample case, the left-out R_ℓ^* , $\ell \neq i, i'$ would not be informative once R_i^* and $R_{i'}^*$ are taken into account.

2.2.3 The case with more than three samples

Theorem 2 states that the joint distribution of any two linear rank statistic is asymptotically bivariate normal. This result together with the Cramer-Wold device yields the joint asymptotic multivariate normality of any $k - 1$ linear rank statistic. Thus, the case with $k > 3$ is handled by considering the correlation matrix for an arbitrary $k - 1$ of the R_i 's in a similar fashion. The test statistic is formed by the exponent of the $(k - 1)$ -variate normal distribution after multiplication by -2 and follows an

asymptotic $\chi^2(k-1)$ distribution under the null hypothesis. An R-code that implements the generalized Kruskal–Wallis test for any $k \geq 2$ is available at author’s website.

2.3 Further Considerations

2.3.1 H -test as a special case

The generalized Kruskal–Wallis test reduces to the original Kruskal–Wallis test when G_j is known for each subject. In this case, we define $p_{ij} = (Z_{1j}, Z_{2j}, \dots, Z_{kj})$. Then, $R_i^* = R_i$ and it is easy to show that

$$\tilde{\mu}_i = \frac{N+1}{2}n_i = E(R_i), \quad \tilde{\sigma}_i^2 = \frac{n}{12}(N+1)(N-n) = \text{Var}(R_i),$$

and that

$$\tilde{\rho}_{ii'} = -\sqrt{\left(\frac{n_i}{N-n_i}\right)\left(\frac{n_{i'}}{N-n_{i'}}\right)} = \text{Cor}(R_i, R_{i'}), \quad \text{for } i \neq i'.$$

Hence, H -test is a special case of H^* .

Remark. *A major advantage of the generalized Kruskal–Wallis test is that it does not require a separate treatment for partially uncertain categorical data. In line with the above approach, when G_j is available, the ranks are allocated to the corresponding groups with probability 1, i.e. $p_{ij} = (Z_{1j}, Z_{2j}, \dots, Z_{kj})$ and when there is uncertainty in the group membership, the probability-weighting principle in (4) guard the testing procedure against possible misclassifications.*

2.3.2 Correction for ties

The generalized Kruskal–Wallis test can be corrected for ties in the same way as the original Kruskal–Wallis test. When ties occur in the data, one typically assigns the mean of the tied ranks to each member of the tied group. This specification, while not affecting the mean rank, reduces the population variance by $\sum T/(12N)$ and increases the population covariance by $\sum T/\{12N(N-1)\}$, where $T = (t-1)t(t+1)$ for each group of ties, t denotes the number of ties in the group, and the summation

is taken over all groups. Thus, in the case of ties, the variance of R_i^* decreases by

$$\frac{\sum T}{12(N-1)} \sum_{j=1}^N (p_{ij} - \bar{p}_i)^2.$$

Similarly, for $i \neq i'$, the covariance between R_i^* and $R_{i'}^*$ is reduced by

$$\frac{\sum T}{12(N-1)} \sum_{j=1}^N (p_{ij} - \bar{p}_i)(p_{i'j} - \bar{p}_{i'}).$$

The corrected generalized Kruskal–Wallis test for ties is hence obtained by substituting the adjusted $\tilde{\sigma}_i$ and $\tilde{\rho}_{ii'}$ in H^* . Note that, in many situations the difference between the generalized Kruskal–Wallis test and its tie-corrected version would be negligible barring an excessive number of ties. The R-code provided at author’s website accommodates tie-correction in the generalized Kruskal–Wallis test.

2.3.3 Exact distribution

The exact null distribution of the Kruskal–Wallis test for three samples, each with up to five observations, is given in [Kruskal and Wallis \(1952\)](#). More extensive tables are later provided by [Iman et al. \(1975\)](#) for up to eight observations in each sample. With a moderately large number of observations, the exact probability calculations of the H statistic become cumbersome, even with modern computing power. In the case of the generalized Kruskal–Wallis test, similar tabulations, even for small samples, seem intractable due to probability weights and remain an open problem. Similar to the suggestion of [Kruskal and Wallis \(1952\)](#), we recommend using the chi-square approximation when the $\sum_{j=1}^N p_{ij}$ ’s are at least five.

2.3.4 Variation in group probabilities

The proposed generalized Kruskal–Wallis test is conditional on observed data. Therefore, in our proposal, we treat the \mathbf{p}_j ’s as fixed quantities that define the underlying probability distribution of the unknown G_j ’s. The treatment is reasonable for genetic applications as there is little variation in the genotype probabilities if the same algorithm is employed more than once provided the input data are the same (e.g. [Pei et al., 2008](#)).

In an unconditional inference, the variation in the group probabilities due to random sampling needs to be taken into account. This amounts to specifying a probability distribution for the p_{ij} 's, which may not be feasible in practice. Here, we briefly discuss the validity of the proposed test when the group probabilities are random but treated as fixed, focusing on the case $k = 2$. For the distribution of the p_{ij} 's, while one may consider, for instance, a mixture of two beta distributions, we prefer to keep our argument as general as possible.

Suppose the vector of probabilities $\{p_{i1}, \dots, p_{iN}\}$ are drawn independently from a probability distribution with mean and variance satisfying conditions analogous to (A1)-(A3). Let μ_i^* and σ_i^{*2} denote the unconditional mean and variance of R_i^* , respectively. It is easy to see that

$$\mu_i^* = \frac{N+1}{2} \sum_{j=1}^N E(p_{ij}) = E(\tilde{\mu}_i),$$

and, by variance decomposition, we obtain

$$\begin{aligned} \sigma_i^{*2} &= \frac{N(N+1)}{12} \sum_{j=1}^N E\{(p_{ij} - \bar{p}_i)^2\} + \left(\frac{N+1}{2}\right)^2 \sum_{j=1}^N \text{Var}(p_{ij}) \\ &= E(\tilde{\sigma}_i^2) + \text{Var}(\tilde{\mu}_i), \end{aligned}$$

where $\tilde{\mu}_i$ and $\tilde{\sigma}_i^2$ remain the conditional mean and the conditional variance of R_i^* .

The statistic in H^* , normalized according to the conditional quantities, can be written as

$$\frac{R_i^* - \tilde{\mu}_i}{\tilde{\sigma}_i} = \frac{\sigma_i^*}{\tilde{\sigma}_i} \left(\frac{R_i^* - \mu_i^*}{\sigma_i^*} - \frac{\tilde{\mu}_i - \mu_i^*}{\sigma_i^*} \right). \quad (7)$$

The first quantity in the parenthesis can be shown to be asymptotically standard normal under certain conditions similar to (A1)-(A3), and the second quantity has an approximate normal distribution with mean zero and variance $\text{Var}(\tilde{\mu}_i)/\sigma_i^{*2}$, by the central limit theorem. Note that

$$\text{Cov}(R_i^*, \tilde{\mu}_i) = \left(\frac{N+1}{2}\right)^2 \sum_{j=1}^N \text{Var}(p_{ij}) = \text{Var}(\tilde{\mu}_i)$$

and hence

$$\frac{R_i^* - \tilde{\mu}_i}{\sigma_i^*} \xrightarrow{\mathcal{L}} N\left(0, \frac{E(\tilde{\sigma}_i^2)}{\sigma_i^{*2}}\right).$$

Using Slutsky’s theorem, we can obtain $(\tilde{R}_i - \tilde{\mu}_i) / \tilde{\sigma}_i \xrightarrow{\mathcal{L}} N(0, 1)$, provided that $\tilde{\sigma}_i^2 \xrightarrow{P} E(\tilde{\sigma}_i^2)$. The latter condition is satisfied when $\text{Var}\{(p_{ij} - \pi_i)^2\} = \tau^2 > 0$.

Thus, under certain assumptions, the generalized Kruskal–Wallis test statistic remains valid when group probabilities are random. The application in Section 4 provides further evidence of this conclusion.

3 Simulations

Here we evaluate the methods via simulation studies. Specifically, (i) we evaluate the validity of the asymptotic null distribution of the generalized Kruskal–Wallis test in finite samples, and (ii) we compare the finite sample performance of the generalized Kruskal–Wallis test with those of commonly used parametric tests as summarized in Table 2. We focus on the dosage approach as it is the most popular method used in practice, and previous work that compared various parametric approaches also recommended its usage (Zheng et al., 2011).

Table 2. Summary of the association tests compared

Test		df	Incorporate Group Uncertainty	Robust to Model Assumptions
<u>I: (p_{0j}, p_{1j}, p_{2j}) is used to determine the most likely genotype group</u>				
BG-LM	Best-Guess Linear Model	1	No	No
BG-ANOVA	Best-Guess ANOVA	2	No	No
BG-KW	Best-Guess Kruskal–Wallis	2	No	Yes
<u>II: (p_{0j}, p_{1j}, p_{2j}) is used directly in the test</u>				
Dosage	Dosage	1	Yes	No
GKW	Generalized Kruskal–Wallis	2	Yes	Yes

3.1 Simulation Methods

As a proof of principle and being consistent with the motivation of this work, we simulated genetic association data. Phenotype and SNP genotype data for $n = 1,000$

individuals were generated as follows.

The three genotype groups were coded as $G = 0, 1$ and 2 copies of the minor allele of a SNP. The SNP of interest had a minor allele frequency of 20%, leading to expected group size of $(n_0, n_1, n_2) = (640, 320, 40)$ based on the multinomial distribution with parameters $(0.64, 0.32, 0.04)$ under the Hardy–Weinberg equilibrium assumption. We also considered minor allele frequency of 10% and other values.

The phenotype data were generated from an additive normal model, favorable to the parametric methods. Without loss of generality, Y values were simulated from normal distribution with equal mean of $(2, 2, 2)$ under the null model, and $(1.75, 2, 2.25)$ under the alternative model for the three genotype groups $G = 0, 1$ and 2, respectively, with a common variance, $\sigma^2 = 1$. The mean values were chosen such that power (at the $\alpha = 0.01$ level) to detect association between Y and G is about 95% for minor allele frequency of 20% (and about 70% for minor allele frequency of 10%) with the given sample size and without genotype uncertainty.

Other simulating parameter values were also considered with varying minor allele frequencies, type 1 error rates and sample sizes, as well as non-normal or non-additive models (Table S1). Additional results are provided in the Supplemental Material and conclusions are characteristically similar to the ones reported here.

Given a true genotype G , to simulate the probabilistic genotype data, we used the Dirichlet distribution with scale parameters a for the correct genotype category and $(1 - a)/2$ for the other two, where $a = 1, 0.9, 0.8$, and 0.7 , corresponding to an increasing level of group uncertainty ranging from 0% to 30%. Under this Dirichlet simulating model, the proportion of the individuals whose most probable (best-guessed) genotypes are the correct ones is approximately a (Table 3).

For each set of probabilistic data, $M = 10,000$ experiments under the null model and 5,000 experiments under the alternative model were conducted by simulating only the response data. Application in Section 4 confirms that methods comparison is not affected by how the probabilistic data were generated.

3.2 Evaluation of Accuracy

Table 4 provides the empirical type 1 error rates of the five tests considered. The group uncertainty does not seem to alter the accuracy of any of the tests in an obvious

Table 3. The empirical proportion of the individuals whose most probable (best-guessed) genotypes are the correct ones. a is the Dirichlet parameter for the correct genotype category. SNP has a minor allele frequency of 20%. Results for other frequencies are similar.

a	average	$G = 0$	$G = 1$	$G = 2$
1	1.00	1.00	1.00	1.00
0.9	0.93	0.92	0.96	0.91
0.8	0.83	0.82	0.85	0.86
0.7	0.74	0.74	0.74	0.80

way. For the proposed GKW test, we also compare the quantiles of the empirical distributions with those of the $\chi^2(2)$ distribution. Figures S1 and S2 indicate that the empirical null distribution coincides with the asymptotic one, which is further supported by the Kolmogorov–Smirnov test with p -values 0.279, 0.595, 0.628 and 0.599 for SNP with minor allele frequency of 20% and 0.174, 0.377, 0.375 and 0.194 for SNP with minor allele frequency of 10%, from the lowest to the highest uncertainty levels.

Table 4. Empirical type 1 error rates of the five tests at $\alpha = 0.01$ under a normal null model, for testing the association of a SNP that has minor allele frequency of 20% or 10%. a is the parameter value of the Dirichlet distribution used to simulate genotype probabilities.

uncertainty level	minor allele frequency = 0.2				minor allele frequency = 0.1			
	0%	10 %	20 %	30 %	0%	10 %	20 %	30 %
	($a = 1$)	($a = 0.9$)	($a = 0.8$)	($a = 0.7$)	($a = 1$)	($a = 0.9$)	($a = 0.8$)	($a = 0.7$)
BG-LM	0.0109	0.0108	0.0090	0.0096	0.0093	0.0091	0.0095	0.0105
BG-ANOVA	0.0098	0.0106	0.0109	0.0094	0.0083	0.0089	0.0103	0.0102
BG-KW	0.0094	0.0091	0.0100	0.0097	0.0075	0.0096	0.0098	0.0094
Dosage	0.0109	0.0105	0.0091	0.0099	0.0093	0.0092	0.0095	0.0095
GKW	0.0094	0.0087	0.0091	0.0088	0.0075	0.0088	0.0092	0.0095

3.3 Evaluation of Efficiency

We examine the empirical relative efficiency of the other tests compared to the proposed generalized Kruskal–Wallis test under the alternative model, using the normal additive data favorable to the model-based tests (Figure 1). The empirical power of each test was obtained using the corresponding empirical type 1 error threshold reported in Table 4. Note that, when the true genotypes are used (i.e. $a = 1$ with 0% group uncertainty), the GKW and the BG-KW are equivalent. This is also true for the dosage test and the BG-LM.

As expected, the dosage has the best power when there is no group uncertainty (Figure 1 at 0% uncertainty level), because the data were generated under the best scenario for the dosage test (i.e. phenotype Y was normally distributed with population means, $(1.75, 2, 2.25)$, increasing in an additive manner with respect to the number of copies of the minor allele, $G = (0, 1, 2)$). When the minor allele frequency is 20%, the power of the dosage test remains (slightly) higher than the generalized Kruskal-Wallis test even as the uncertainty level increases (Figure 1(a)). However, this is no longer true for detecting SNPs with minor allele frequency of 10% (Figure 1(b)). For example, with uncertainty level at 30% ($a = 0.7$), the relative efficiency of all other tests including the dosage test is about 60% as compared to the generalized Kruskal-Wallis test.

Moreover, in less favorable models, (e.g. Figure S7 for non-normal model), the generalized Kruskal-Wallis test can outperform the others even when there is no genotype uncertainty. Additional simulation results with different minor allele frequencies (e.g. 0.05 and 0.3), different type 1 error rate (e.g. 0.05 and 0.001) and different model assumptions (e.g. non-additive model) as presented in Figures S4-S8 all confirm the robustness of the generalized Kruskal-Wallis test. Under scenarios favorable to model-based methods, the generalized Kruskal-Wallis test provides comparable power; with model misspecification or increased genotype uncertainty, the generalized Kruskal-Wallis test can outperform the others.

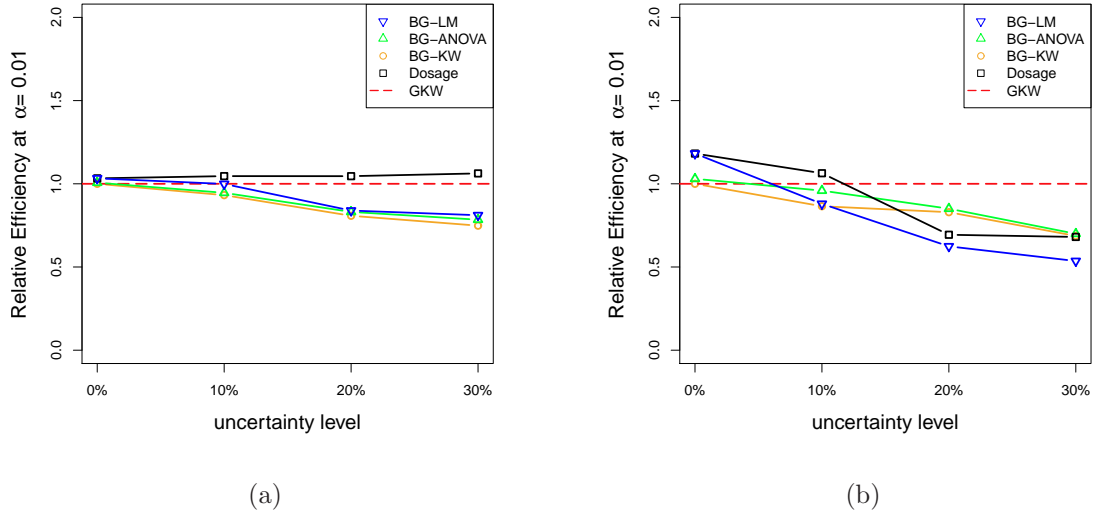


Figure 1. Relative efficiency of other tests as compared to the GKW test, under a normal additive model, for testing the association of a SNP that has (a) minor allele frequency of 20%, or (b) minor allele frequency of 10%.

4 Data Applications

We demonstrate the proposed method in three applications using data from the genome-wide association study of complications in type 1 diabetes patients, for which over 800K SNPs were genotyped by the Illumina 1M beadchip assay and over 1.5 million ungenotyped SNPs were imputed (Paterson et al., 2010).

The study sample consists of $n = 1,300$ subjects with type 1 diabetes (664 treated conventionally and 636 treated intensively) from the Diabetes Control and Complications Trial (DCCT). The phenotypes of interest are glycosylated hemoglobin (HbA1c) and diastolic blood pressure (DBP), collected quarterly from each patient over the course of the DCCT.

The first application illustrates the case of no association using 27,265 ungenotyped but imputed SNPs on chromosome 22. The other two applications evaluate the performance of the generalized Kruskal–Wallis test in detecting putative associations of two genotyped SNPs with DBP and HbA1c.

4.1 P -value Distribution of H^* on Chromosome 22

We investigated the p -value distribution of the generalized Kruskal–Wallis test using chromosome 22 data under the null hypothesis of no association. The genotype data at 33,815 SNPs were imputed (provided by Dr. Andrew Paterson’s research group) using MaCH (Li et al., 2006, 2009) using HapMap II phased data as the reference panel. In the association analysis, we considered 27,265 SNPs that yielded the sum of genotype probabilities of at least five for each of the three genotype groups.

The phenotype data, mean DBP measurements over the first six study periods, were first permuted to eliminate any possible associations. We observed that the null distribution of p -values obtained from the $\chi^2(2)$ approximation of the generalized Kruskal–Wallis test statistic closely matches the theoretical uniform distribution. Figure 2 displays the corresponding quantile-quantile plot on the $-\log_{10} p$ scale. The Kolmogorov–Smirnov test for uniformity yields p -value= 0.084, providing additional evidence for the validity of the proposed generalized Kruskal–Wallis test in finite samples, because in contrast to the simulations, the group probabilities are random in that their distributions vary between imputed SNPs and among the patient subjects.

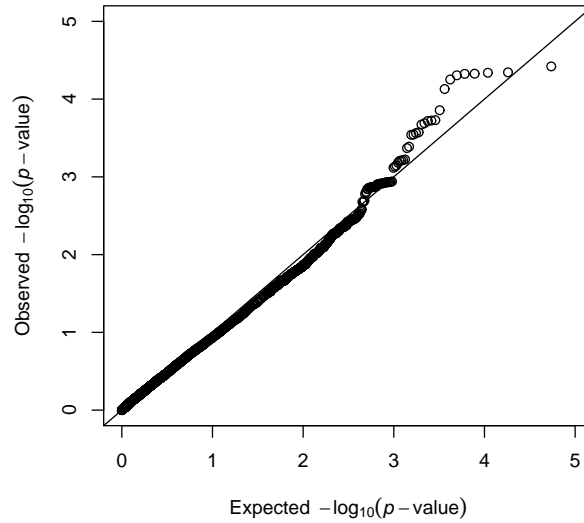


Figure 2. The quantile-quantile plot of the $-\log_{10}(p\text{-values})$ of the GKW test applied to assess association between 27,265 imputed SNPs on chromosome 22 and permuted diastolic blood pressure phenotype measures in 1,300 subjects with type 1 diabetes.

4.2 DBP with rs7842868

Recently, [Ye et al. \(2010\)](#) identified rs7842868 on chromosome 8 as a SNP associated with DBP with p -value $\approx 4.5 \times 10^{-8}$. Here, we examined the performance of the five tests in detecting this association. Since GKW and BG-KW are readily advantaged in the case of non-normal data, we base our comparisons on the natural logarithm of the DBP measurements averaged over the first six study periods. The histogram of the phenotype data for the 1,300 patients is given in Figure 3(a).

The observed genotype data at rs7842868 yielded the group sizes $(n_0, n_1, n_2) = (788, 446, 66)$. When tested for association, rs7842868 was found to be significant by all tests with similar results (Figure 3(a) at 0% uncertainty level). GKW (equivalent to BG-KW when genotypes are known) had a marginally higher statistical significance with p -value $= 1.36 \times 10^{-8}$, followed by the dosage test and BG-LM with p -value $= 2.66 \times 10^{-8}$.

We then masked the actual data and simulated 1,000 replicates of *in silico* genotypes (i.e. genotype probabilities) from the Dirichlet distribution at each group uncertainty level, as described in Section 3. The results averaged over the 1,000 replications are shown in Figure 3(a) and verify the robustness of the GKW test. For instance, at 10% or higher uncertainty level, the proposed GKW test had noticeable better performance than all other procedures. Nevertheless, power to detect association by any method deteriorated considerably as genotype uncertainty increased.

4.3 HbA1c with rs1358030

The SNP rs1358030 on chromosome 10 was reported to be genome-wide significantly associated with HbA1c (p -value $= 5 \times 10^{-9}$) in the conventionally treated group ([Paterson et al., 2010](#)). We performed association analyses using the observed genotypes and masked probabilistic genotypes in a fashion similar to the above DBP application.

The histogram of the phenotype, average $\log(\text{HbA1c})$ measurements over the first six study periods, is given in Figure 3(b). for the $n = 664$ conventionally treated patients. The genotype group sizes at rs1358030 were $(n_0, n_1, n_2) = (267, 307, 90)$. In this application, when the actual genotypes were used, the dosage test (equivalent to BG-LM) showed the best performance in detecting the HbA1c association (Figure 3(b) at 0% uncertainty level). However, the advantage of the dosage test dissipated

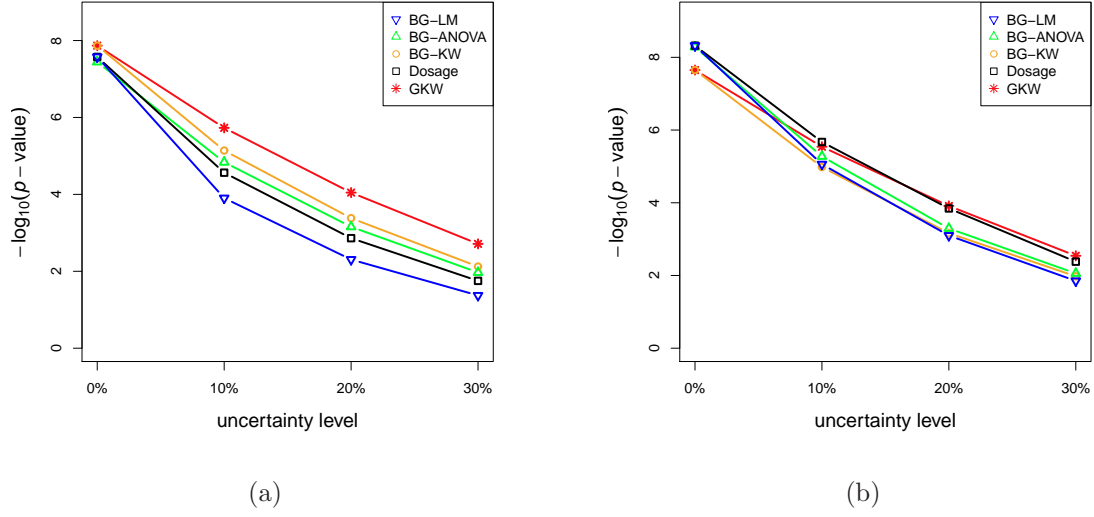


Figure 3. The significance, on the $-\log_{10}(p\text{-value})$ scale, of the association tests at different genotype uncertainty levels for (a) DBP with rs7842868, (b) HbA1c with rs1358030.

as the genotype uncertainty level increased.

5 Conclusions and Disucssions

In this paper, we generalized the rank-based nonparametric Kruskal–Wallis test to allow for group uncertainty when comparing k samples. The proposed generalized test statistic follows an asymptotic chi-square distribution with $k - 1$ degrees of freedom, suitable for statistical inference of large-scale data (e.g. genome-wide association or next-generation sequencing data) without the need for permutation or other computationally inefficient procedures.

Although the work was originally motivated by the analyses of SNPs with genotype uncertainty in genetic association studies, it can be readily applied to other scientific studies. Extensive simulations and several applications showed that the generalized Kruskal–Wallis test provides a good balance between robustness and power. Our proof-of-principle stimulation studies could be improved to more closely mimic real genetic data and models. However, the validity and robustness conclusion characteristically holds, given the combined evidence from our theoretical work, and simulation

and application studies.

The proposed generalized Kruskal–Wallis test has its limitations. For example, the exact distribution for small samples (e.g. relevant to genetic association studies of rare variants) is unknown. The current test does not include other covariates. However, this limitation could be partially alleviated by considering the residuals from a regression model that accounts for the effects of other covariates first, assuming there is no SNP-covariate interaction. Both the original and the generalized Kruskal–Wallis tests are formulated for continuous outcomes, however, the probability-weighting principle exploited here could potentially be extended to analyze case-control data or other categorical outcomes and is the subject of ongoing research.

In conclusion, the proposed generalized Kruskal–Wallis test provides scientists a tool to continue investigate, in a robust non-parametric fashion, the k -sample problems in the presence of group uncertainty. The generalized Kruskal–Wallis test includes the original Kruskal–Wallis test as a special case, and its power is comparable to parametric counterparts under conditions favorable to the model-based approaches. When there is model misspecification or high group uncertainty, the generalized Kruskal–Wallis test can outperform the others.

Acknowledgement

The authors thank Dr. Andrew Paterson and his research group, specifically Daryl Waggott and Ye Chang, for providing their genome-wide association data and Drs. Fang Yao and D.A.S. Fraser for constructive discussions. The research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC, 250053-2008) and Canadian Institutes of Health Research (CIHR, MOP 84287).

References

- Aulchenko, Y. S., Struchalin, M. V., and van Duijn, C. M. (2010), “ProbABEL package for genome-wide association analysis of imputed data,” *BMC Bioinformatics*, 11.
- Calvo, S. E., Tucker, E. J., Compton, A. G., Kirby, D. M., Crawford, G., Burt, N. P.,

- et al. (2010), “High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency.” *Nature Genetics*.
- Carvalho, B. S., Louis, T. A., and A, I. R. (2010), “Quantifying uncertainty in genotype calls,” *Bioinformatics*, 26, 242–249.
- Fraser, D. A. S. (1957), *Nonparametric methods in statistics*, New York: Wiley.
- Hájek, J., Sidák, Z., and Sen, P. K. (1999), *Theory of Rank Tests*, Academic Press.
- Iman, R. L., Quade, D., and Alexander, D. A. (1975), “Exact probability levels for the Kruskal-Wallis test,” in *Selected tables in mathematical statistics*, eds. Harter, H. L. and Owen, D. B., American Mathematical Society, vol. 3, pp. 329–384.
- Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., et al. (2008), “Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs,” *Nature Genetics*, 40, 1253–1260.
- Kruskal, W. H. (1952), “A nonparametric test for the several sample problem,” *Annals of Mathematical Statistics*, 23, 525–540.
- Kruskal, W. H. and Wallis, W. A. (1952), “Use of ranks in one-criterion variance analysis,” *Journal of American Statistical Association*, 47, 583–621.
- Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., et al. (2010), “Methods for testing association between uncertain genotypes and quantitative traits,” *Biostatistics*.
- Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2006), “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes,” *Genetic Epidemiology*, 34, 816–834.
- Li, Y., Willer, C. J., Sanna, S., and Abecasis, G. R. (2009), “Genotype Imputation,” *Annual Review of Genomics and Human Genetics*, 10, 387–406.
- Lin, D. Y., Hu, Y., and Huang, B. E. (2008), “Simple and efficient analysis of disease association with missing genotype data,” *American Journal of Human Genetics*, 83, 535–539.

- Marchini, J. and Howie, B. (2010), “Genotype imputation for genome-wide association studies,” *Nature Review Genetics*, 11, 499–511.
- Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007), “A new multipoint method for genome-wide association studies by imputation of genotypes,” *Nature Genetics*, 39, 906–913.
- Nicolae, D. L. (2006), “Testing Untyped Alleles (TUNA)—Applications to Genome-Wide Association Studies,” *Genetic Epidemiology*, 30, 718–727.
- Paterson, A. D., Waggott, D., Borigt, A. P., Hosseini, S. M., Shen, E., Sylvestre, M.-P., et al. (2010), “A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose,” *Diabetes*, 59, 539–49.
- Pei, Y.-F., Li, J., Zhang, L., Papasian, C. J., and H-W, D. (2008), “Analyses and Comparison of Accuracy of Different Genotype Imputation Methods,” *PLoS ONE*, 3, e3551.
- Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., and Poland, G. A. (2002), “Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous,” *American Journal of Human Genetics*, 70, 425–434.
- Wald, A. and Wolfowitz, J. (1944), “Statistical Tests Based on Permutations of the Observations,” *Annals of Mathematical Statistics*, 15, 358–372.
- Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011), “SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data.” *Nucleic Acids Research*, 39.
- Ye, C., Canty, A. J., Waggott, D., Sylvestre, M.-P., Shen, E., Hosseini, M., et al. (2010), “A Repeated Measures Genome Wide Association Study of Blood Pressure in Type 1 Diabetes.” Abstract # 203 presented at the Nineteenth Annual Meeting of the International Genetic Epidemiology Society. *Genetic Epidemiology*, 34, 973.
- Zheng, J., Li, Y., Abecasis, G. R., and Scheet, P. (2011), “A Comparison of Approaches to Account for Uncertainty in Analysis of Imputed Genotypes,” *Genetic Epidemiology*, 35, 102–110.

Supplemental Material

This supplement contains the results of additional simulations under various scenarios, as well as the tables and figures cited in the main text. Table S1 summarizes the study purpose and the settings of our simulation studies where sample size is 1,000. Other sample sizes (e.g. 500 or 2,000) were also studied, but results were categorically similar, therefore not reported here.

Table S1. Reader Guide for the simulation results.

Study the impact of	Power	Normal	Additive	α	MAF
minor allele frequency	Figure 1(a)	yes	yes	0.01	0.2
	Figure 1(b)	yes	yes	0.01	0.1
minor allele frequency	Figure S4(a)	yes	yes	0.01	0.05
	Figure S4(b)	yes	yes	0.01	0.3
type 1 error rate	Figure S5(a)	yes	yes	0.05	0.2
	Figure S5(b)	yes	yes	0.001	0.2
	Figure S6(a)	yes	yes	0.05	0.1
	Figure S6(b)	yes	yes	0.001	0.1
model assumptions	Figure S7(a)	no	yes	0.01	0.2
	Figure S7(b)	no	yes	0.01	0.1
	Figure S8(a)	yes	no	0.01	0.2
	Figure S8(b)	yes	no	0.01	0.1

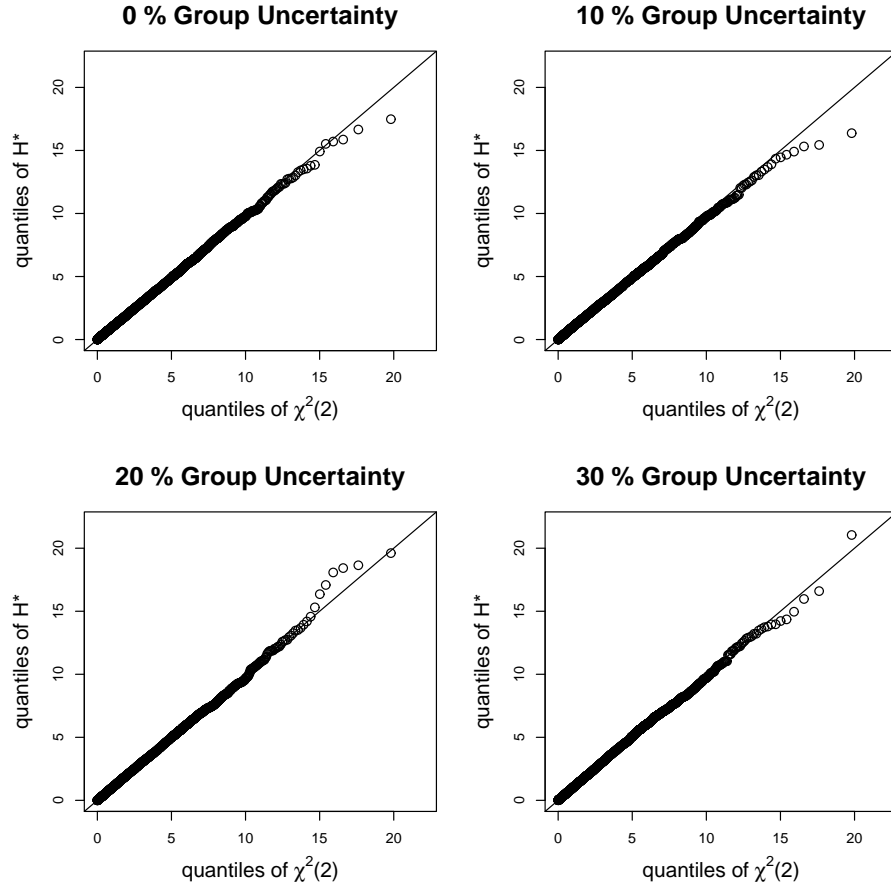


Figure S1. The quantile-quantile plots for the GKW test statistic, under a normal null model, for testing the association of a SNP that has minor allele frequency of 20%. The Kolmogorov–Smirnov test has p-values of 0.279, 0.595, 0.628 and 0.599, for the lowest to the highest uncertainty levels.

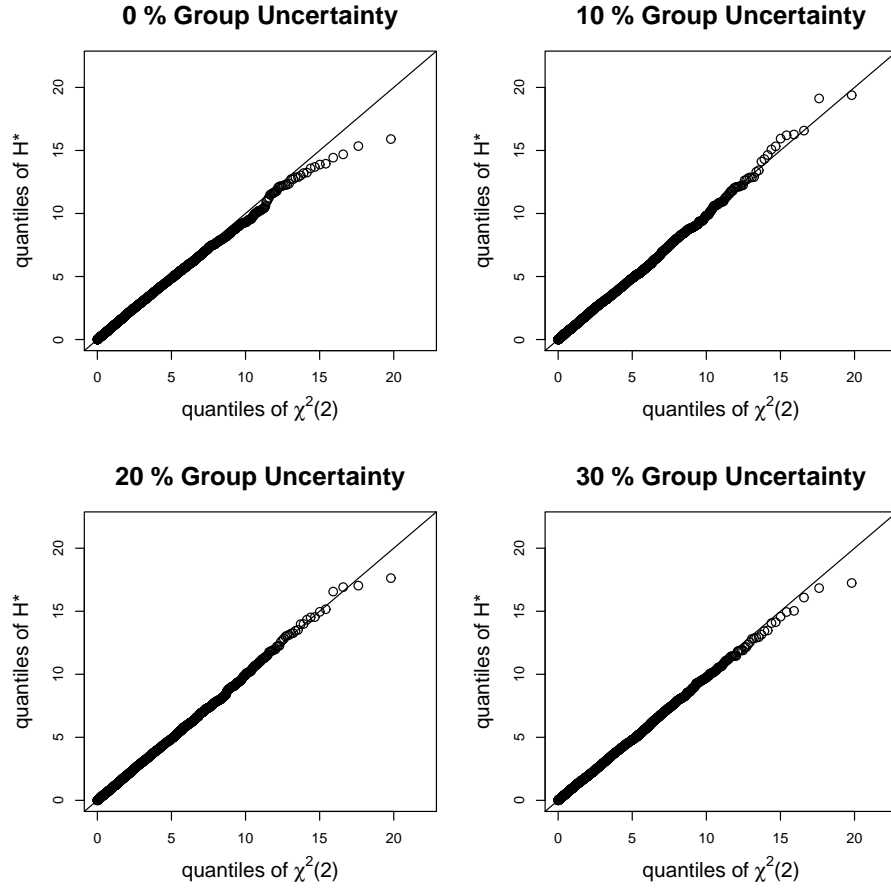
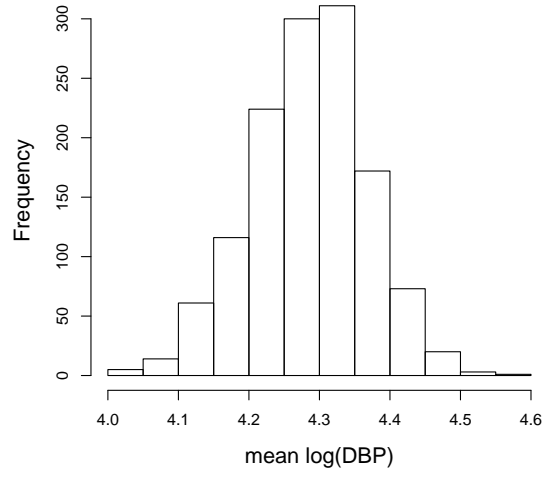
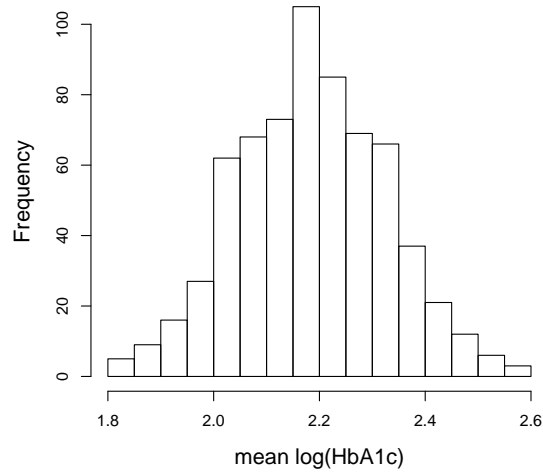


Figure S2. The quantile-quantile plots for the GKW test statistic, under a normal null model, for testing the association of a SNP that has minor allele frequency of 10%. The Kolmogorov–Smirnov test has p-values of 0.174, 0.377, 0.375 and 0.194, for the lowest to the highest uncertainty levels.

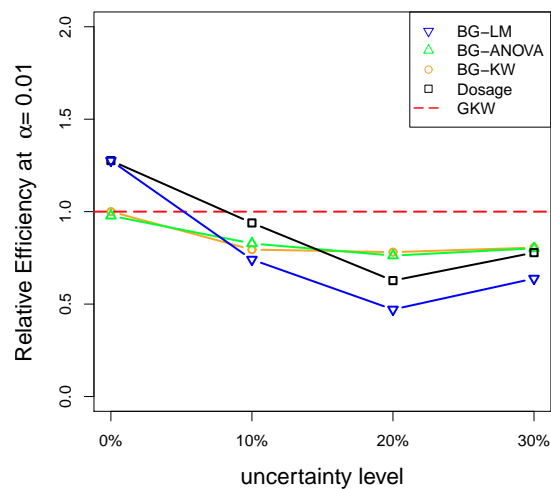


(a)

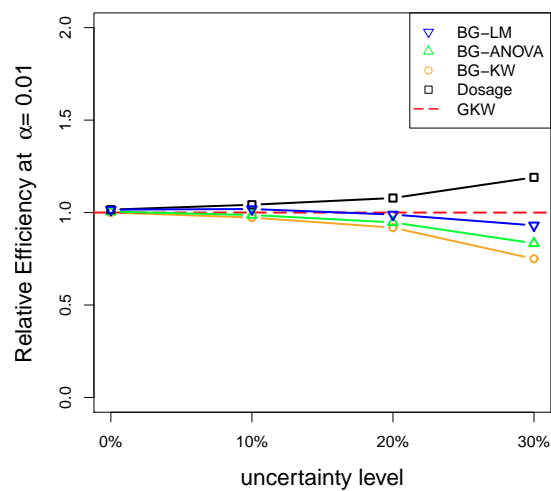


(b)

Figure S3. Histograms of (a) mean of the natural logarithm of DBP measurements of 1,300 patients over the first six study periods, (b) mean of the natural logarithm of HbA1c measurements of 664 conventionally treated patients over the first six study periods.

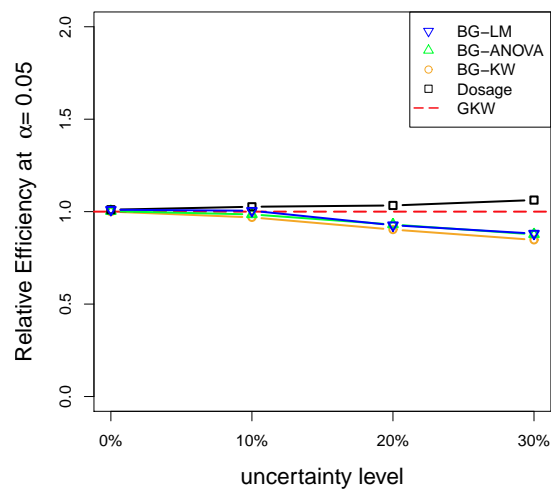


(a)

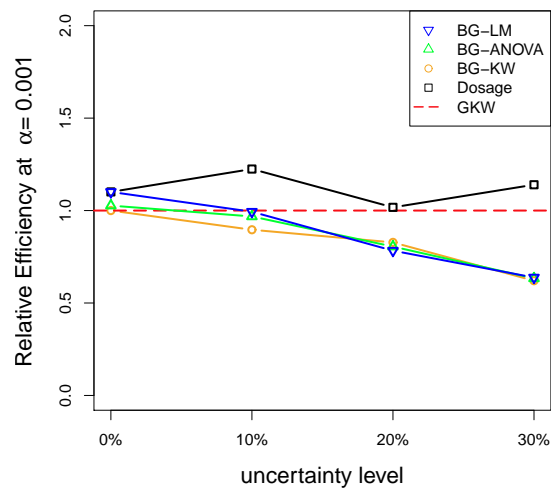


(b)

Figure S4. Relative efficiency of other tests as compared to the GKW test at $\alpha = 0.01$, under a normal additive model, for testing the association of SNP that has (a) minor allele frequency of 5%, (b) minor allele frequency of 30%.

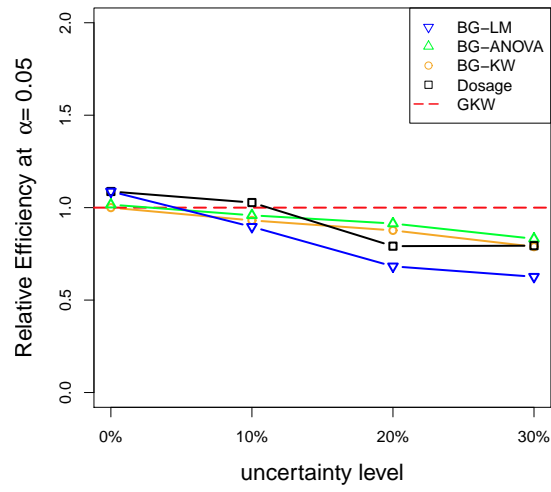


(a)

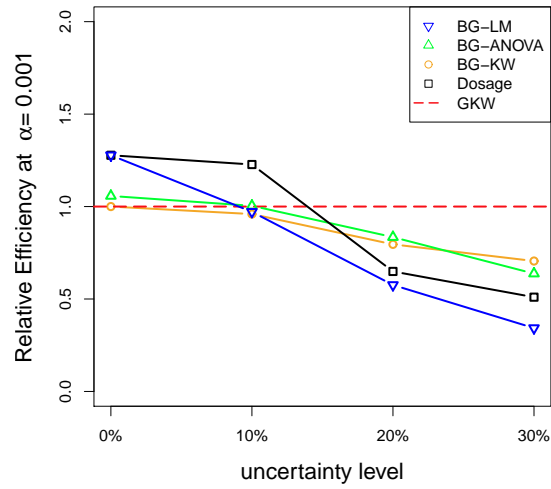


(b)

Figure S5. Relative efficiency of other tests as compared to the GKW test at (a) $\alpha = 0.05$, (b) $\alpha = 0.001$, under a normal additive model, for testing the association of SNP that has minor allele frequency of 20%.

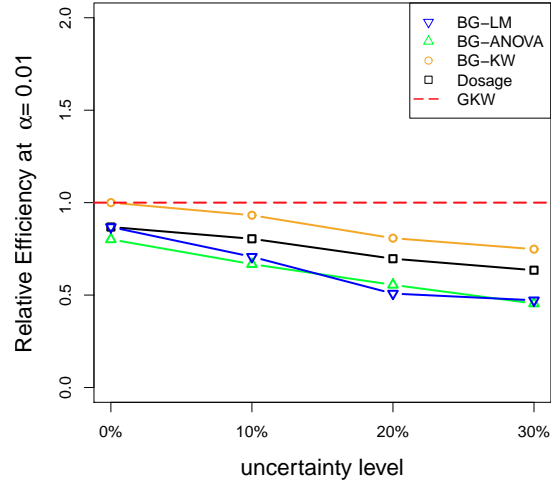


(a)

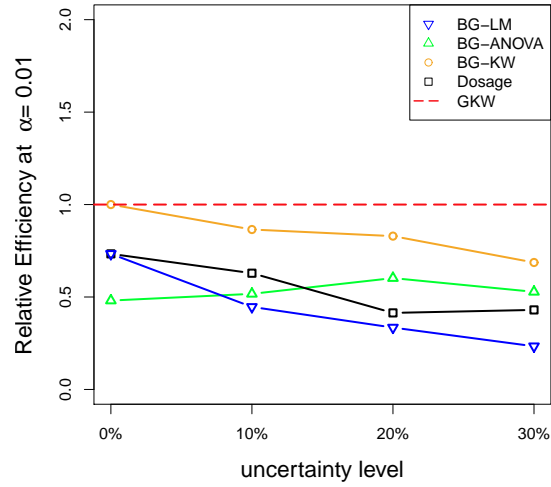


(b)

Figure S6. Relative efficiency of other tests as compared to the GKW test at (a) $\alpha = 0.05$, (b) $\alpha = 0.001$, under a normal additive model, for testing the association of SNP that has minor allele frequency of 10%.

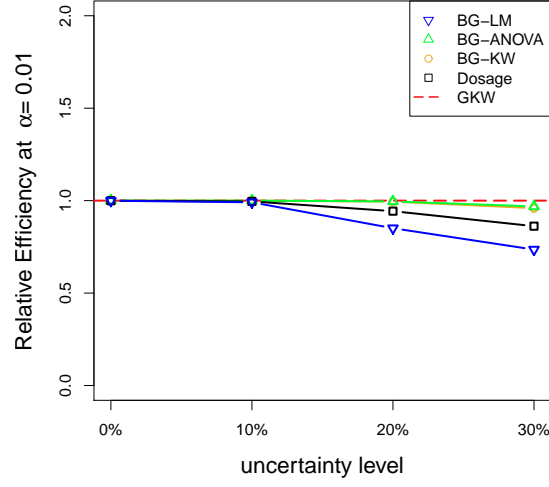


(a)

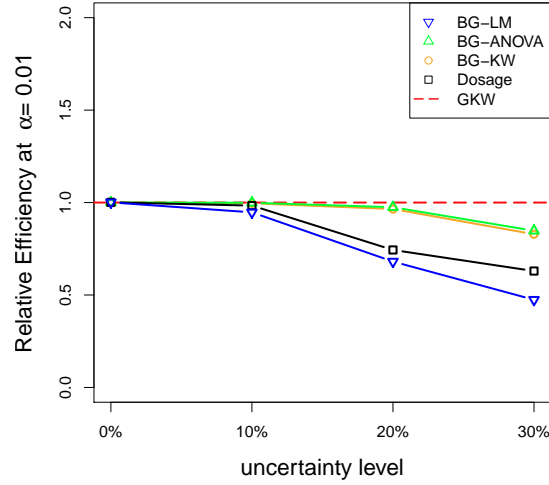


(b)

Figure S7. Relative efficiency of other tests as compared to the GKW test at $\alpha = 0.01$ under a non-normal additive model, for testing the association of SNP that has (a) minor allele frequency of 20%. (b) minor allele frequency of 10%. The non-normal data were obtained by taking the exponent of the normal data generated as in Section 3.



(a)



(b)

Figure S8. Relative efficiency of other tests as compared to the GKW test at $\alpha = 0.01$ under a normal non-additive model, for testing the association of SNP that has (a) minor allele frequency of 20%. (b) minor allele frequency of 10%. The data under non-additive model were generated from normal distribution with means (1.75, 2.25, 2) for the three genotype groups $G = 0, 1$ and 2, respectively, with a common variance, $\sigma^2 = 1$.